ELSEVIER

Information Fusion 1 (2000) 17–33

INFORMATION
FUSION

www.elsevier.nl/locate/inffus

Multi-modal identity verification using expert fusion

Patrick Verlinde [a,*], Gérard Chollet [b], Marc Acheroy [a]

[a] Royal Military Academy, Signal and Image Centre, Renaissancelaan 30, B-1000 Brussels, Belgium
[b] ENST/TSI Department, CNRS URA-820, Paris, France

Received 10 February 1999; received in revised form 27 July 1999; accepted 11 November 1999

Abstract

The contribution of this paper is to compare paradigms coming from the classes of parametric, and non-parametric techniques to solve the decision fusion problem encountered in the design of a multi-modal biometrical identity verification system. The multi-modal identity verification system under consideration is built of *d* modalities in parallel, each one delivering as output a scalar number, called score, stating how well the claimed identity is verified. A decision fusion module receiving as input the *d* scores has to take a binary decision: accept or reject the claimed identity. We have solved this fusion problem using parametric and non-parametric classifiers. The performances of all these fusion modules have been evaluated and compared with other approaches on a multi-modal database, containing both vocal and visual biometric modalities. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Multi-modal identity verification; Biometrics; Decision fusion

1. Introduction

The automatic verification [1] of a person is increasingly becoming an important tool in several applications such as controlled access to restricted (physical and virtual) environments. Just think about secure teleshopping, accessing the safe room of your bank, telebanking, or withdrawing money from automatic teller machines (ATMs).

A number of different, readily available techniques, such as passwords, *smart* cards and personal identification numbers (PIN) are already widely used in this context, but the only thing they really verify is, in the best case, a combination of a certain *possession* (for instance the possession of the correct smart card) and of a certain *knowledge*, through the correct restitution of a character and/or digit combination. As is well known, these intrinsically simple (access) control mechanisms can very easily lead to abuses, induced for instance by the loss or theft of the smart card and the corresponding PIN. Therefore a new kind of method is emerging, based on the so-called *biometric* characteristics or measures, such as voice, face (including profile), gait, eye (iris-pattern, retina-scan), fingerprint, hand-shape or some other unique and measurable physiological or behavioral characteristic information of the person to be identified. Biometric measures in general, and non-invasive/user-friendly (vocal, visual) biometric measures in particular, are very attractive because they have the huge advantage that one cannot lose or forget them, and they are really personal (one cannot pass them to someone else), since they are based on a physical appearance measure. We can start using these user-friendly biometric measures now, thanks to the progress made in the field of automatic speech analysis and artificial vision. In this paper, the term *modality* is reserved for a biometric measure. An *expert* is each algorithm or method using characteristic features coming from a particular biometric measure to verify the identity of a person under test.

If one uses only a single (user-friendly) biometric measure, the results obtained may be found to be not good enough. This is due to the fact that these user-friendly biometric measures tend to *vary with time* for one and the same person and to make it even worse, the importance of this variation is itself very variable from one person to another. This is especially true for the vocal (speech) modality, which shows an important *intra-speaker variability*. One possible solution to try to cope with the problem of this *intra-person* variability is *to use more than one* (*user-friendly*) *biometric measure*. In this new *multi-modal* context, it is thus becoming

[*] Corresponding author. Tel.: +32-2-7376621; fax: +32-2-7376622.
E-mail addresses: verlinde@elec.rma.ac.be (P. Verlinde), chollet@tsi.enst.fr (G. Chollet), acheroy@elec.rma.ac.be (M. Acheroy).

[1] Verification is the binary process of accepting or rejecting the identity claim made by the person under test.

1566-2535/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S 1 5 6 6 - 2 5 3 5 (0 0) 0 0 0 0 2 - 6

important to be able to combine (*or fuse*) the outcomes of different modalities and thus also of different experts. There is currently a significant international interest in this topic. The organization of already two international conferences on the specific subject of *audio- and video-based person authentication* (*AVBPA*) is probably the best proof of this [3,10].

Some work on multi-modal biometric identity verification systems has already been reported in the literature. As early as 1993, Chibelushi et al. [11] have proposed to integrate acoustic and visual speech (motion of visible articulators) for speaker recognition, using a simple linear combination scheme. Brunelli and Falavigna [9] have proposed a person identification system based on acoustic and visual features, where they use a HyperBF network as the best performing fusion module. Dieckmann et al. [17] have proposed a decision level fusion scheme, based on a *2-out-of-3* majority voting, which integrates two biometric modalities (face and voice), analyzed by three different experts: (static) face, (dynamic) lip motion, and (dynamic) voice. Duc et al. [19] did propose a simple averaging technique and compared it with the Bayesian integration scheme presented by Bigün et al. [2]. In this multi-modal system the authors use a face identification expert, and a text-dependent speech expert. Jourlin et al. [22] have proposed an acoustic–labial speaker verification method. Their approach is based on a lip tracker using visual features, and on a text-dependent speech expert. The fused score is computed as the weighted sum of the scores generated by the two experts. Kittler et al. [24] have proposed a multi-modal person verification system, using three experts: frontal face, face profile, and voice. The best combination results are obtained for a simple sum rule. Hong and Jain [20] have proposed a multi-modal personal identification system which integrates two different biometrics (face and fingerprints) that complement each other. The fusion algorithm operates at the expert (*soft*) decision level, where it combines the scores from the different experts (under the statistically independence hypothesis), by simply multiplying them. Ben-Yacoub [1] did propose a multi-modal data fusion approach for person authentication, based on support vector machines (SVM) to combine the results obtained from a face identification expert, and a text-dependent speech expert. Pigeon [32] proposed a multi-modal person authentication approach based on simple fusion algorithms to combine the results coming from three experts: frontal face, face profile, and voice. Choudhury et al. [13] did propose a multi-modal person recognition system using unconstrained audio and video. The combination of the two experts is performed using a Bayes net.

In all these studies, the primary attention is always focused on the choice and the implementation of the different biometrics. The main contribution of this paper is to put the emphasis on the fusion module, and to compare a large number of fusion paradigms for this multi-modal biometric identity verification application.

2. Characterization of an identity verification system

In this paper, we will consider the verification of the identity of a person as a typical two-class problem: either the person is the one (in this case he is called a *client*), or is not the one (in that case he is called an *impostor*) he claims to be. This means that we are going to work with a binary {accept, reject} decision scheme.

When dealing with binary hypothesis testing, it is trivial to understand that the decision module can make two kinds of errors. Applied to this problem of the verification of the identity of a person, these two errors are called

- False rejection (FR): i.e. when an actual *client* is rejected as being an *impostor*:
- False acceptance (FA): i.e. when an actual *impostor* is accepted as being a *client*.

The performances of a speaker verification system are usually given in terms of the global error rates computed during tests: the false rejection rate (FRR) and the false acceptance rate (FAR) [4]. These error rates are defined as follows:

$$\mathrm{FRR} = \frac{\text{number of FRs}}{\text{number of client accesses}}, \tag{1}$$

$$\mathrm{FAR} = \frac{\text{number of FAs}}{\text{number of impostor accesses}}. \tag{2}$$

A perfect identity verification (FAR = 0 and FRR = 0) is in practice unachievable. However, as shown by the study of binary hypothesis testing [42], any of the two FAR, FRR can be reduced to an arbitrary small value by changing the decision threshold, with the drawback of increasing the other one. A unique measure can be obtained by combining these two errors into the total error rate (TER) or its complimentary, the total success rate (TSR)

$$\mathrm{TER} = \frac{\text{number of FA} + \text{number of FR}}{\text{total number of accesses}}, \tag{3}$$

$$\mathrm{TSR} = 1 - \mathrm{TER}. \tag{4}$$

However, care should be taken when using one of these two unique measures. Indeed, from the definition just given it follows directly that these two unique numbers could be heavily biased by one or either type of errors (FAR or FRR), depending solely on the number of accesses that have been used in obtaining these respective errors. As a matter of fact, due to the proportional weighting as specified in the definition, the TER will always be closer to that type of error (FAR or FRR) which has been obtained using the largest number of accesses.

The overall performance of an identity verification system is however better characterized by its so-called *receiver operating characteristic* (*ROC*), which represents the FAR as a function of the FRR [42]. The detection error trade-off (DET) curve is a convenient nonlinear transformation of the ROC curve, which has become the standard method for comparing performances of speaker verification methods used in the annual NIST evaluation campaigns [35]. In a DET curve, the horizontal axis shows the normal deviate of the False Alarm probability in (%), which is a nonlinear transformation of the horizontal False Acceptance axis of the classical ROC curve. The vertical axis of the DET curve represents normal deviate of the Miss probability (in %), which is a nonlinear transformation of the FR axis of the classical ROC curve. The use of the normal deviate scale moves the curves away from the lower left when performance is high, making comparisons between different systems easier. It can also be observed that, typically, the resulting curves are approximately straight lines, which do correspond to normal likelihood distributions, for at least a wide portion of their range. Further details of this nonlinear transformation are presented in [27]. Figs. 1 and 2 give, respectively, an example of a typical ROC and a typical DET curve.

Each point on a ROC or a DET characteristic corresponds with a particular decision threshold. The equal error rate (EER: i.e. when FAR = FRR), is often used as the only performance measure of an identity verification method, although this measure gives just one point of the ROC and comparing different systems solely based on this single number can be very misleading [31].

High security access applications are concerned about break-ins and hence operate at a point on the ROC with small FAR. Forensic applications desire to catch a criminal even at the expense of examining a large number of false accepts and hence operate at small FRR/high FAR. Civilian applications attempt to operate at the operating points with both low FRR and low FAR. These concepts are shown in Fig. 3, which was found in [21].

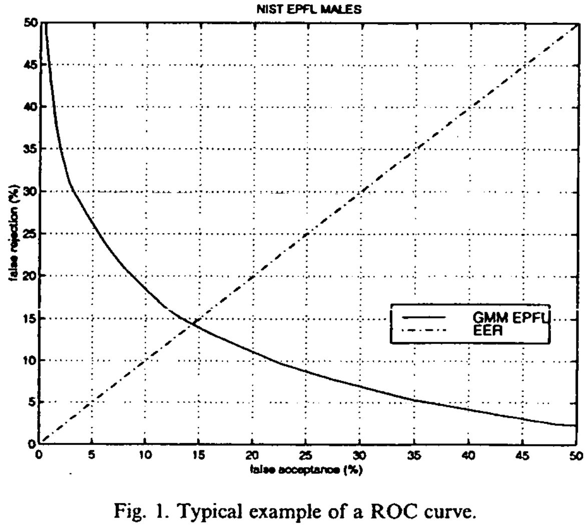

Fig. 1. Typical example of a ROC curve.

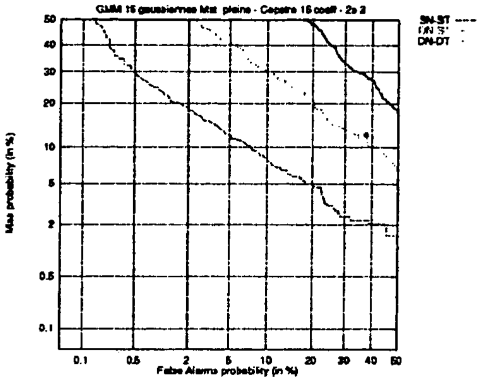

Fig. 2. Typical example of a DET curve.

Unfortunately in practice, as will be shown further in the study of the fusion modules presented in this thesis, it is not always possible to explicitly identify a continuous decision threshold in a certain fusion module, which means that in that case it will a fortiori not be possible to vary the decision threshold to obtain a ROC or a DET curve. So in these specific cases only a single operating point on the ROC can be given. This is incidentally also the only correct way of determining the performance of an operational system, since in such systems the decision threshold has been *fixed.*

All verification results in this thesis will be given in terms of FRR, FAR, and TER. For each error the 95%

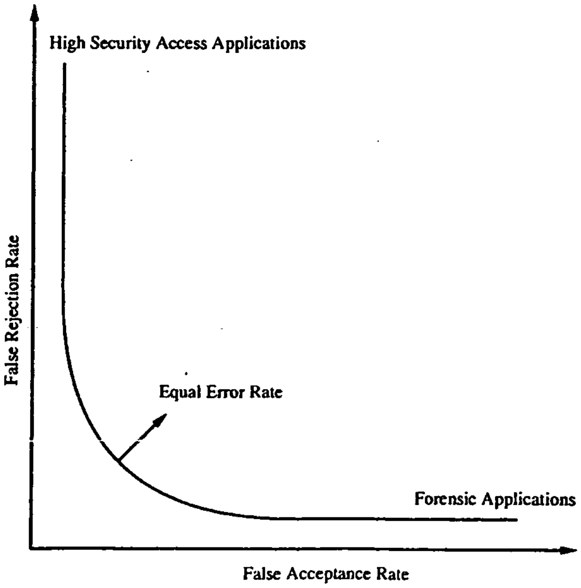

Fig. 3. Typical examples of different operating points for different application types.

level confidence interval will be given between square brackets. The concept of *confidence intervals* refers to the inherent uncertainty in test results owing to small sample size. These intervals are a posteriori estimates of the uncertainty in the results on the test population. They do not include the uncertainties caused by errors (mislabeled data, for example) in the test process. The confidence intervals do not represent a priori estimates of performance in different applications or with different populations [48].

These confidence levels will be calculated assuming that the probability distribution for the number of errors is binomial. But since the binomial law cannot be easily analyzed in an analytical way, the calculation of confidence intervals cannot be done directly in an analytical way. Therefore we have used the Normal law as an approximation of the binomial law. This large sample approach is already statistically justified starting from 30 samples. Using this approximation, the 95% confidence interval of an error E based on N tests, is defined by the following lower (given by the minus sign) and upper (given by the plus sign) bounds:

$$E \pm 1.96\sqrt{\frac{E(1-E)}{N}}.$$

More detailed information about the calculation of confidence intervals can be found in [12,14,37].

3. Experimental protocol

All tests have been carried out using the multi-modal M2VTS database [33]. In this protocol we use the four sessions of the M2VTS database in the following manner:

1. The first enrollment session has been used for training the individual experts. This means that each access has been used to model the respective client, yielding 37 different client models.
2. Then the accesses from each person in the second enrollment session have been used to generate validation data in two different manners. Once to derive one single client access by matching the shot of a specific person with its own reference model, and once to generate 36 impostor access by matching it to the 36 models of the other persons of the database. This simple strategy thus leads to 37 client and $36 \times 37 = 1.332$ impostor accesses, which have been used for validating the performance of the individual experts and for calculating thresholds.
3. The third enrollment session has been used to test these experts, using the thresholds calculated on the validation data set. This same data set has also been used to train the fusion modules, which again leads to 37 client and 1.332 impostor reference points.
4. Finally, the fourth enrollment session has been used to test the fusion modules, yielding once more the same number of client and impostor claims.

The drawback of this simple protocol, is that the impostors are *known* at the expert and supervisor training time. In Section 7.2, validation results will be presented using a protocol that does not suffer from the same drawback. This validation protocol is implemented using a so-called *leave-one-out* method [16].

4. Identity verification experts

4.1. Short presentation

All the experiments in this thesis have been performed using three different identity verification experts. Each one of these experts will be described briefly hereafter.

4.1.1. Profile image expert

The profile image verification expert is described in detail in [34] and its description hereafter has been inspired by the presentation of this expert in [24]. This particular profile image expert is based on a comparison of a candidate profile of the person under test with the template profile corresponding to the claimed identity. The candidate image profile is extracted from the profile images by means of color-based segmentation. The similarity of the two profiles is measured using the Chamfer distance computed sequentially [8]. The efficiency of the verification process is aided by pre-computing a distance map for each reference profile. The map stores the distance of each pixel in the profile image to the nearest point on the reference profile. As the candidate profile can be subject to translation, rotation and scaling, the objective of the matching stage is to compensate for such geometric transformations. The parameters of the compensating transformation are determined by minimizing the chamfer distance between the template and the transformed candidate profile. The optimization is carried out using a simplex algorithm which requires only the distance function evaluation and no derivatives. The convergence of the simplex algorithm to a local minimum is prevented by a careful initialization of the transformation parameters. The translation parameters are estimated by comparing the position of the nose tip in the two matched profiles. The scale factor is derived from the comparison of the profile heights and the rotation is initially set to zero. Once the optimal set of transformation parameters is determined, the user is accepted or rejected depending on the relationship of the minimal chamfer distance to a prespecified threshold. The system can be trained very easily. It is sufficient to store one profile per client in the training set.

4.1.2. Frontal image expert

The frontal image verification expert is described in detail in [28] and the description hereafter was based on the presentation of this expert in [24]. This frontal image expert is based on robust correlation of a frontal face image of the person under test and the stored face template corresponding to the claimed identity. A search for the optimum correlation is performed in the space of all valid geometric and photometric transformations of the input image to obtain the best possible match with respect to the template. The geometric transformation includes translation, rotation and scaling, whereas the photometric transformation corrects for a change of the mean level of illumination. The search technique for the optimal transformation parameters is based on random exponential distributions. Accordingly, at each stage the transformation between the test and reference images is perturbed by a random vector drawn from an exponential distribution and the change is accepted if it leads to an improvement of a matching criterion. The score function adopted rewards a large overlap between the transformed face image and the template, and the similarity of the intensity distributions of the two images. The degree of similarity is measured with a robust kernel. This ensures that gross errors due to, for instance, hair style changes do not swamp the cumulative error between the matched images. In other words, the matching is benevolent, aiming to find as large areas of the face as possible, supporting a close agreement between the respective gray-level histograms of the two images. The gross errors will be reflected in a reduced overlap between the two images, which is taken into account in the overall matching criterion. The system is trained very easily by means of storing one template for each client. Each reference image is segmented to create a face mask which excludes the background and the torso as these are likely to change over time.

4.1.3. Vocal expert

The vocal identity verification expert is presented in detail in [6]. This text-independent speaker verification expert is based on a similarity measure between speakers, calculated on second order statistics [5].

In this algorithm a first covariance matrix X is generated from a *reference* sequence, consisting of M m-dimensional acoustical vectors, and pronounced by the person who's identity is claimed

$$X = \frac{1}{M}\sum_{i=1}^{M} X_i X_i^{\mathrm{T}},$$

where X_i^{T} is X_i *transposed.*

A second covariance matrix Y is then generated in the same way from a sequence, consisting of M m-dimensional acoustical vectors, and pronounced by the person under test.

Then a similarity measure between these two speakers is performed, based on the *sphericity measure* $\mu_{AH}(X;Y)$

$$\mu_{AH}(X,Y) = \log\frac{A}{H},$$

$$A(\lambda_1,\lambda_2,\ldots,\lambda_m) = \frac{1}{m}\sum_{i=1}^{m}\lambda_i = m^{-1}\,\mathrm{tr}(YX^{1}),$$

$$H(\lambda_1,\lambda_2,\ldots,\lambda_m) = m\left(\sum_{i=1}^{m}\frac{1}{\lambda_i}\right)^{-1} = m(\mathrm{tr}(XY^{-1}))^{-1}.$$

It can be shown that this sphericity measure is always non-negative and it is equal to zero only in the case that the two covariance matrices X and Y are the same. The verification process consists then of comparing the obtained sphericity measure with a decision threshold, calculated on a validation database.

One of the great advantages of this algorithm is that no explicit extraction of the m eigenvalues λ_i is necessary, since the sphericity measure only needs the calculation of the trace $\mathrm{tr}(\cdot)$ of the matrix product YX^{-1} or XY^{-1}.

4.2. Performances

The performances achieved by the three mono-modal identity verification systems which have been used in these experiments are given in Table 1. The results have been obtained by adjusting the threshold at the EER on the validation set and applying this threshold as an a priori threshold on the test set. Observing the results for the profile an the frontal experts it can be seen that, although the optimization has been done according to the EER criterion, the FRR and the FAR are very different. This indicates that for these two experts, the training and validation sets are not very representative of the test set.

Table 1
Verification results for individual experts

Expert	FRR (%) (37 tests)	FAR (%) (1.332 tests)	TER (%) (1.369 tests)
Profile	21.6 [11.4, 37.2]	8.5 [7.1, 10.1]	8.9 [7.5, 10.5]
Frontal	21.6 [11.4, 37.2]	8.3 [6.9, 9.9]	8.7 [7.3, 10.3]
Vocal	5.4 [1.5, 17.7]	3.6 [2.7, 4.7]	3.7 [2.8, 4.8]

4.3. Statistical analysis of the different experts

4.3.1. Introduction

A statistical analysis of the individual experts [2] is important to get an idea on the one hand of their individual discriminatory power, and of their complementarity on the other.

The power of an expert to discriminate between clients and impostors will increase (for given variances) with the difference between the mean value of the scores obtained for client accesses and the mean value of the scores obtained for impostor accesses. The typical statistical test to see if there exist significant differences between the means (or more generally between the statistical moment of first order) of several populations is the so-called analysis of variance (ANOVA). In the general case, this analysis is implemented using an *F*-test. In the specific case of two populations, this ANOVA could also be performed using an independent samples *t*-test [30]. Another important characteristic of an expert is its variance (or more generally the statistical moment of second order). The equality of variances can be tested by a Levene test, which is also implemented using an *F*-test [26]. It is advantageous that the variance of an expert is the same for clients and for impostors, because this leads to simpler methods to combine the different experts. Obviously we will need to perform *t*- and *F*-tests to analyze the means and the variances of the different experts. However, the *t*- and *F*-tests give only exact results if the populations have a Normal distribution. So before we can use *t*- or *F*-tests, we need to verify the Normality of the different populations. Thus this is the first statistical analysis that we need to perform. Since the ANOVA is only valid if the variances of the different populations per expert are equal, we have to check the equality of variances before performing the ANOVA. These remarks explain the forced order of the first three analyses that are presented below.

We can get an idea of the independence of the different experts (and thus of the amount of *extra* information that each expert brings in), by analyzing their correlation. And a linear discriminant analysis gives us a first idea of the combined discriminatory power of the experts.

Last but not least, the analysis of the extreme values gives us insight into the possible use of personalized approaches.

4.3.2. Analysis of Normality

The purpose of a Normality analysis is to check whether the observed data do or do not support the hypothesis (H0) that the underlying probability density function is Normal. There exist two types of tests to perform this analysis: objective (numerical) and subjective (graphical) tests. An important remark related to the verification of H0 is that the assumption of Normality is much more difficult to verify when using small sample sizes. In a sample of small size, the probability of verifying that the data is coming from a Normal distribution is actually very small.

The best known representative of the objective/numerical type of tests is the so-called Kolmogorov–Smirnov (K–S) test for goodness of fit, applied to the Normal distribution [23]. The results obtained by this test on our data are presented in Table 2. This table shows the values obtained for the K–S statistic, the degrees of freedom (df) and the *significance* of this test at the 95% confidence level. This confidence level leads to a critical value for the significance of 0.05. If the significance is smaller than this critical value, then we *reject* the Normality hypothesis H0. If on the other hand the significance is greater than the critical value, then we say that we do not have enough evidence to reject H0, so in a binary decision concept we are forced to *accept* H0 [30].

To be able to analyze the results obtained by this K–S test, it is important to know that its severity increases with the sample size of the population. This means that the K–S test is not severe for small sample sizes (as is the case for the client populations), but very severe for large sample sizes (as is the case for the impostor populations). This means that if the results lead to an acceptance of H0, and if the sample size is sufficiently large, then the Normality assumption is very good. But in the case of a rejection of H0, this does not mean that we cannot accept H0 at all. In that case we need more information to be able to decide, and therefore we have to go on to the second type of Normality tests: the subjective/graphical tests. In our case, the only two populations that are not being rejected by the K–S test as being Normal are the client distributions for the frontal and the vocal experts. But since the sample sizes coming from these two distributions are very small [37], this result has to be used with great care. For all the other populations there is enough evidence to reject the hypothesis H0.

Table 2
Results for the Kolmogorov–Smirnov test for Normality

Population	Statistic	df	Significance
Profile clients	0.227	37	0.000
Profile impostors	0.195	1332	0.000
Frontal clients	0.133	37	0.096
Frontal impostors	0.052	1332	0.000
Vocal clients	0.087	37	0.200
Vocal impostors	0.060	1332	0.000

[2] All the following statistical tests have been performed using the SPSS statistical software package [40].

There exist several types of graphical representations, which can be used as subjective tests. A first useful type of graphical representation is the so-called Normal Q–Q plot [40]. This kind of representation is shown in Fig. 4. The idea is to judge if the plotted sample points do follow *sufficiently* (this is clearly subjective) the ideal line which is the representation of a perfect Normal distribution. Depending on this subjective opinion, we reject or not the Normality assumption.

A second representation is the detrended variant of the first one [40], which is shown in Fig. 5. In this type of graphical representation, one needs to judge whether or not *enough* (another subjective measure) sample points are situated between 2.0 and −2.0 standard normal units. If this is the case wwe accept the Normality assumption. In the other case, we reject it.

Finally, a third representation can be obtained by plotting the histograms of the different populations and by comparing them with the ideal Bormal distribution plots [40]. This kind of graphical representation is given in Fig. 6. The idea is to check *how well* (again subjective) the actual histograms match the ideal Normal distribution plot.

Taking into account the results of the objective K–S test and having inspected these graphical representations, we conclude that the Normality assumptions in the strict sense of the word are not fulfilled for our populations. This is especially true for the profile expert, and in some lesser extent to the vocal expert. The frontal expert deviates the least from a Normal deviation. In practice this means a real drawback, because if the Normality hypothesis is satisfied, this generally leads to substantial simplifications.

However, since the observed deviations of Normality are not too important, and taking into account that the classical *t*- and *F*-tests are robust with respect to

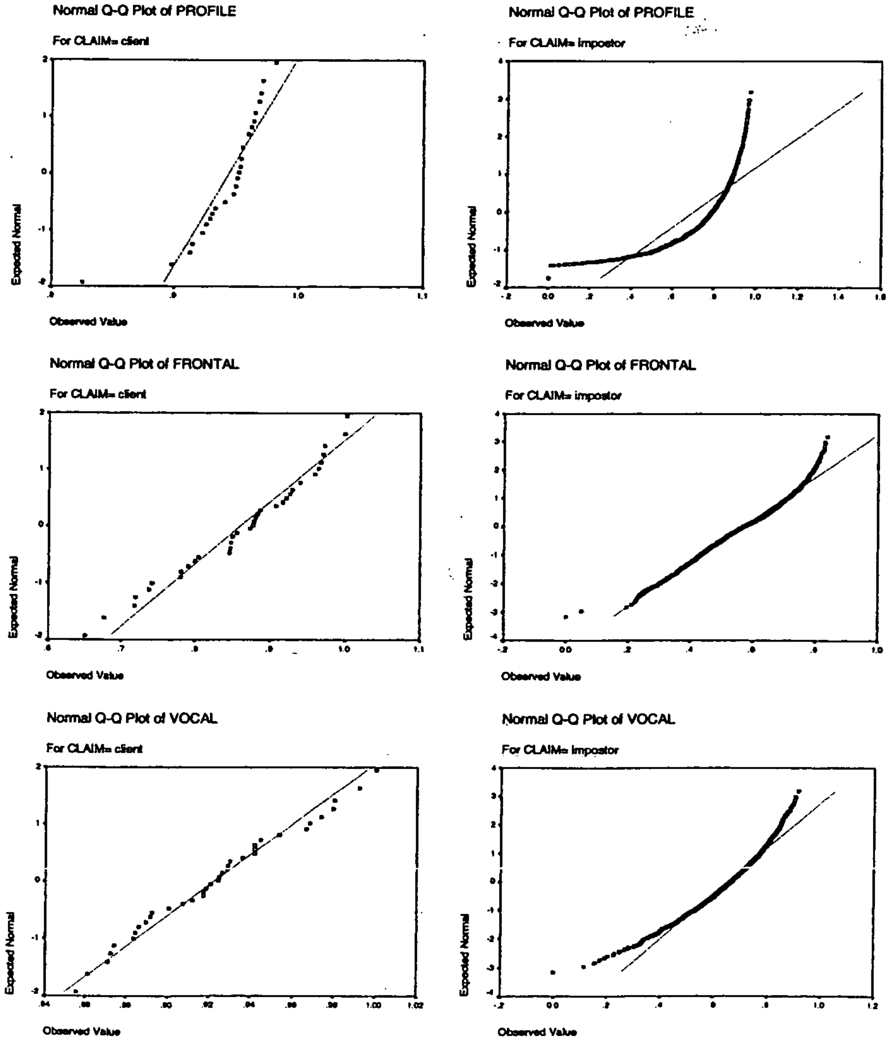

Fig. 4. Normal Q–Q plots for clients and impostors, ranked per expert.

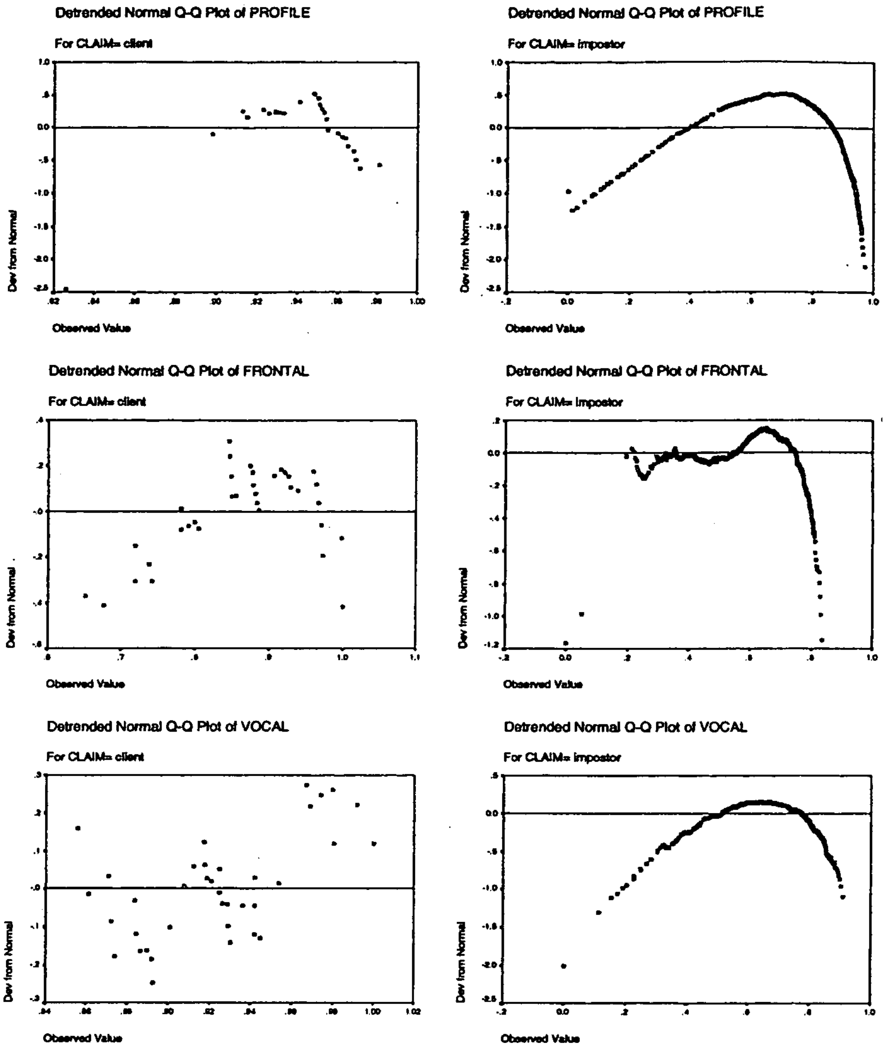

Fig. 5. Detrended Normal Q–Q plots for clients and impostors, ranked per expert.

deviations from Normality, we are nevertheless going to perform *t*- and *F*-tests, but with the important restriction that the results of these tests will have to be analyzed with the utmost care. In other words, we will accept the results obtained by *t*- and *F*-tests if and only if they have a significance level which is far away from the critical value (0.05 for the 95% confidence interval).

4.3.3. Analysis of variance

The results obtained by the Levene test are given in Table 3. The H0 hypothesis is that there are no differences in the variances of the different populations. As we can see by the significance of 0.000 for the vocal and the profile experts and of 0.003 for the frontal expert, this H0 hypothesis is strongly rejected for all experts. Since the rejection is so strong, we do accept the results of this Levene test and conclude that all experts have significant different variances for both populations.

To confirm the results of this analysis, we can have a look at a *box-plot* representation [3] of the different experts, presented in Fig. 7. From these representations it follows that for each expert the variance of the client population is indeed significantly smaller than the variance of the impostor population.

[3] The 'box' in the box-plot is delimited at the bottom by the 25th and at the top by the 75th percentile. The height of the box thus gives an idea of the variance. The black line in the middle of the box represents the median (50th percentile), which is a robust estimation of the mean. The whiskers underneath and on top of the box, respectively, show the lowest and the highest values, with the exception of outliers (represented by a circle) and extreme values (represented by an asterisk). An *outlier* is defined by a value which is situated 1.5 times the thickness of the box outside the box, and an *extreme value* is defined by a value which is situated 3 times the thickness of the box outside the box [40].

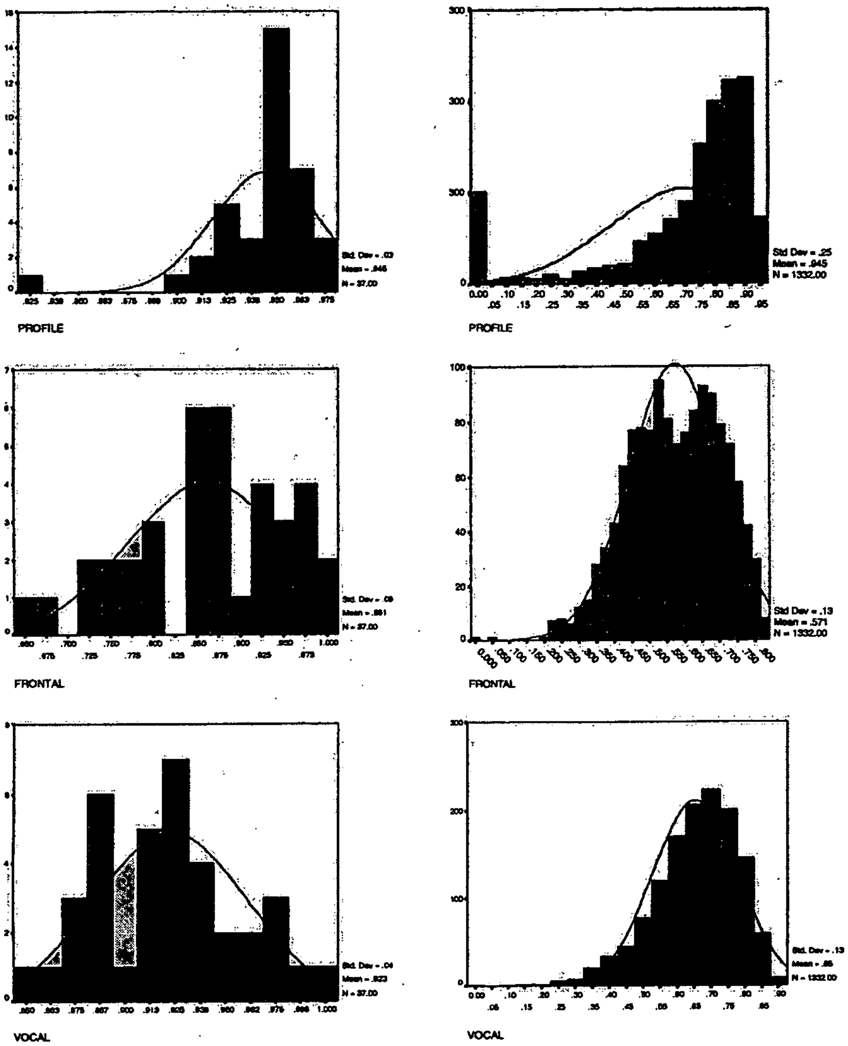

Fig. 6. Histogram plots for clients and impostors, showing the Normal distribution, and ranked per expert.

Table 3
Results for the Levene test for equality of variances

Population	F-Statistic	Significance
Profile	33.125	0.000
Frontal	9.062	0.003
Vocal	28.284	0.000

Since an ANOVA for analyzing the differences between the means of the populations can, strictly speaking, only be used with Normal distributions and equal variances, we do have a problem here. Fortunately, since we have only two different populations, we can also use an independent samples t-test, which can be calculated as well for equal variances (in this case the t-test is in an exact form) as for unequal variances (this time the t-test is in an approached form). This means that we will calculate the difference of means hereafter, using an independent samples t-test with *unequal* variances.

4.3.4. Analysis of means

The results of the independent samples t-test with unequal variances are given in Table 4. The H0 hypothesis is that there are no differences between the means of the different populations. As we can see by the significance of 0.000, this H0 hypothesis is strongly rejected for all experts. Since the rejection is so strong, we do accept the results of this t-test and conclude that all experts have significant different means for both populations.

To confirm the results of this analysis, we can again have a look at the box-plot representation of the different experts, presented in Fig. 7. In these representations it can be seen that for each expert the median of the client population is indeed significantly higher than

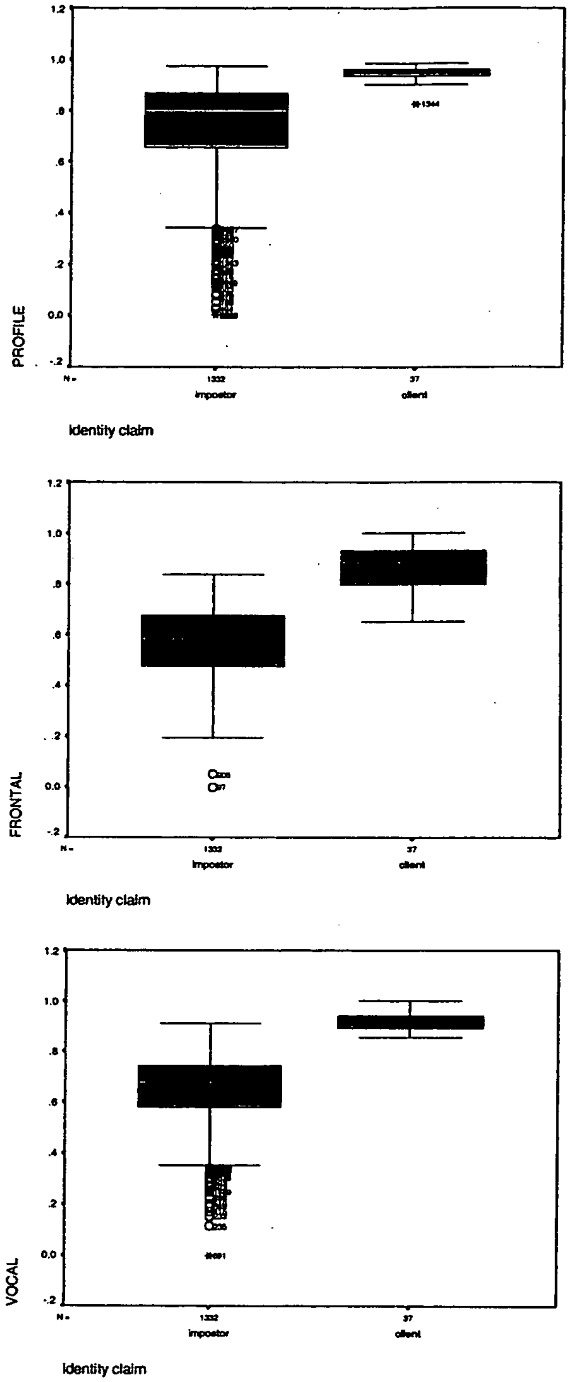

Fig. 7. Box-plots giving for each expert an idea of the means and variances for the client and impostor scores.

Table 4
Results for the independent samples t-test with unequal variances for detecting differences in means

Expert	t-Statistic	df	Significance
Profile	−29.398	372.665	0.000
Frontal	−18.855	40.275	0.000
Vocal	−38.198	61.642	0.000

the median of the impostor population. And since the median is a robust estimation of the mean, the same conclusions are valid for the mean.

4.3.5. Analysis of correlation

Another important statistical analysis is the calculation of the correlation that exists between the different experts. A popular way of seeing the importance of this is to say that the more the errors the different experts make are de-correlated, the better our fusion could get since the amount of *new* information introduced by each expert will tend to be larger. The correlation matrix is represented in Table 5. As could be expected from the diversity of the experts we are using, the correlation is very low.

This correlation can also be visualized by taking the experts two-by-two and plotting the observed results for each population. These matrix scatter plots are shown in Fig. 8, and the fact that all the distributions have shapes

Table 5
Correlation matrix for our three experts

Correlation	Profile	Frontal	Vocal
Profile	1.000	0.011	−0.043
Frontal	0.011	1.000	0.256
Vocal	−0.043	0.256	1.000

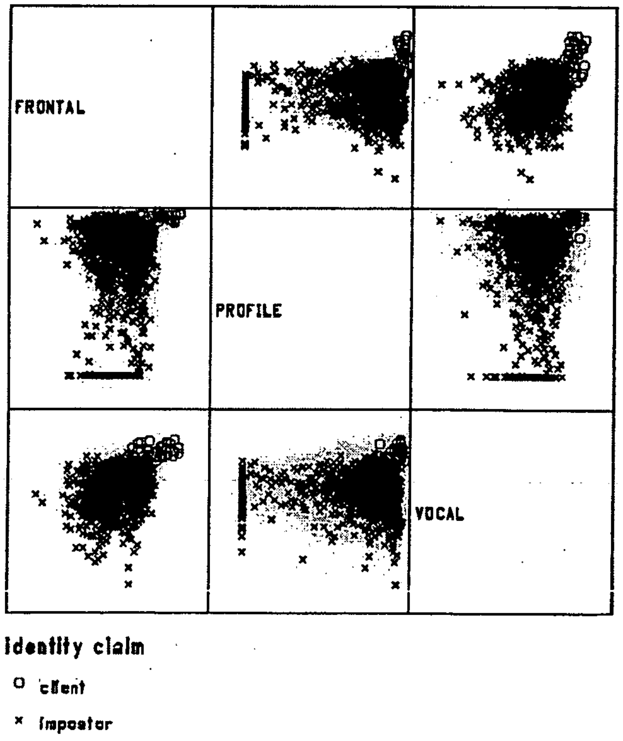

Fig. 8. Visual representation of the correlation of the different experts taken two-by-two.

in the form of a cloud indicates that the correlation is very low.

A direct conclusion from this correlation analysis is that a principal component analysis (PCA) is not useful here, because of the low correlation between the different experts. Another reason is obviously the fact that we only have used three experts, so there is no real need for performing a data reduction by means of PCA.

4.3.6. Linear discriminant analysis

To examine the discriminatory power and the complementarity of the three experts at the same time, we have performed a linear discriminant analysis [26,40]. The results of this linear discriminant analysis are shown in Table 6.

When we compare the results of the linear discriminant analysis with those obtained by the individual modalities, it is clear that combining the three experts does lead to far better performances, even if the combination is done using a simple linear classifier. This indicates that the different experts do have enough discriminatory power and are sufficiently complementary to make it worthwhile to investigate the combination problem in more detail.

4.3.7. Analysis of extreme values

Another important point in descriptive statistics is the analysis and the handling of extreme values. Normally speaking, extreme values should be discarded from the calculation of characteristic statistical measures such as means or variances. In our work however we will not do that, since these extreme values can contain interesting information as will be shown hereafter.

Indeed, in Fig. 7 it can be seen that the profile expert presents a large number of extreme values (represented by asterisks) for impostor accesses. These extreme values can also be observed in Fig. 8, where they form very specific alignments in the four sub-plots where the scores of the profile expert are plotted against one of the axis.

One of the possible explanations for this phenomenon, taking into account the experimental protocol, is that the profile of one the clients is very different from the profiles of all other clients. After analyzing the scores of the profile expert it turned out that, when claiming the identity of client number eight, 22 out of the 36 other clients obtained a score equal to zero! This phenomenon is represented under form of matrix scatter plots in Fig. 9.

Table 6
Results of the linear discriminant analysis (LDA)

Method	FRR (%) (37 tests)	FAR (%) (1332 tests)	TER (%) (1369 tests)
LDA	0.0 [0.0, 9.4]	5.4 [4.3, 6.7]	5.3 [4.2, 6.6]

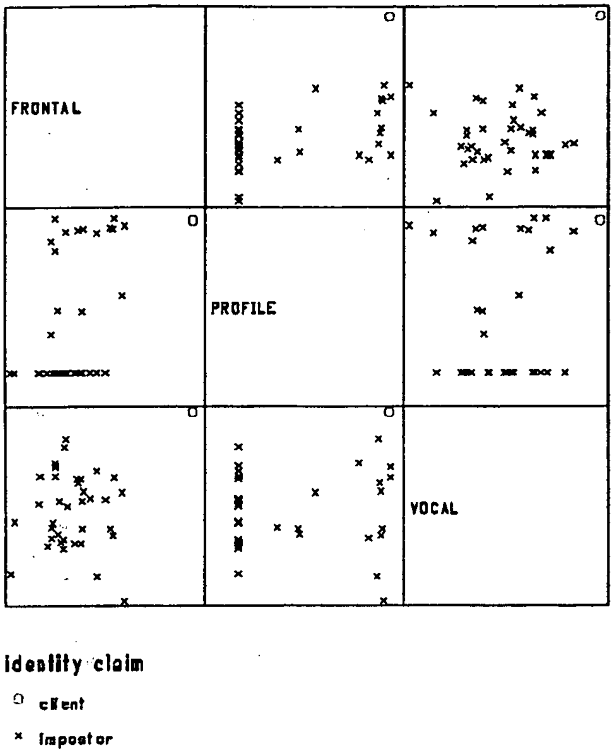

Fig. 9. Matrix plots representing the scores for all 37 persons claiming the identity of client number eight (i.e. one client access and 36 impostor accesses).

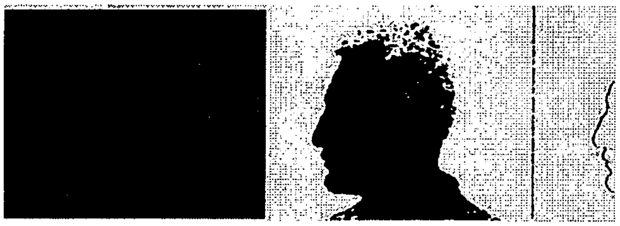

Fig. 10. Typical profile images of client number 8.

In order to understand this phenomenon, it is interesting to have a look at some typical profile images of client number eight. Such images are presented in Fig. 10.

From these profile images, it is easy to see that the chin of this specific client is very pronounced, which makes it a very personal characteristic. Therefore few profiles of other clients of the database will present good results when being matched against this typical profile, which obviously leads to very good impostor rejection performances. This observation suggests that, in some specific cases, a *personalized approach* based for instance on specific characteristics of certain persons, might improve system performance substantially. This is an interesting observation, especially when seen in the light of the actual efforts to come to *robust* methods, in which extreme values, such as the ones we have been considering here, are very likely to be excluded!

In this work we did however not use such a personalized approach, since the chosen application does not provide enough training data. This will also become obvious in Section 7.2, where it will be shown that the validation protocol which is presented there, does not support a personalized approach.

5. Decision fusion in identity verification

Combining the partial decisions from the d different experts in a decision fusion strategy without considering the temporal fusion aspect, could be done using one of the two following basic architectures [15]:

Serial suite. As shown in Fig. 11, a *serial* expert architecture consists of a set of d *experts* whose decisions are combined in series or tandem. This architecture is for instance well-suited to deal with situations where the different experts do not use a binary {accept, reject}, but rather a ternary {accept, reject, undecided} decision scheme. If in the latter case, the current expert cannot decide, he hands over the information he has on to the next expert in the sequence. For this serial scenario to be effective, the next expert in line obviously needs to be designed as a real *expert* in dealing with the cases that cannot be solved by the previous expert. This architecture is thus particularly well-suited to combine the decisions from experts which have varying ranges of effectiveness and to model sequential decision refining from one sensor to the next. This is not the case in our problem.

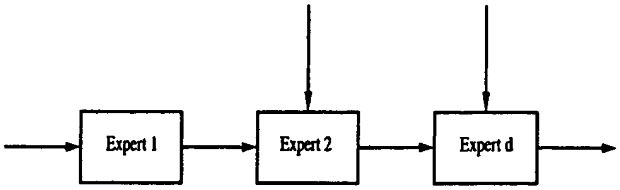

Fig. 11. A typical serial multi-expert decision fusion architecture.

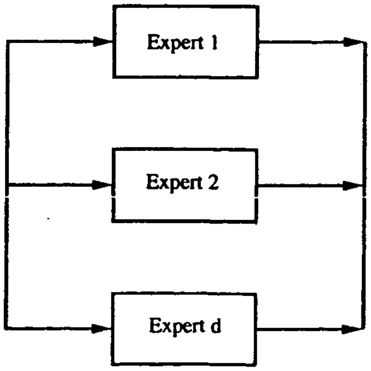

Fig. 12. A typical parallel multi-expert decision fusion architecture.

Parallel suite. As shown in Fig. 12, a *parallel* expert architecture consists of a set of d experts that are interrogated in parallel. The decisions derived from these experts are combined in parallel by the fusion module. This architecture is particularly well-suited to combine the decisions or scores from experts that are capable of operating simultaneously and independently of one another. This is the case in our problem.

Next to the two fairly simple architectures presented above, one can also imagine more complicated combinations of these two basic schemes, such as parallel–serial or serial–parallel architectures. These combinations are more complex than the previous two and fall outside the scope of this work. Another possible extension of architectures presented so far, is the introduction of some kind of generalized feed-back mechanism. In this case, the idea is to postpone the decision until for instance a new set of measurements has been taken. The basic idea behind this technique can be illustrated by the following example: if a vocal expert is undecided in a ternary decision scheme, the automatic verification system could prompt the user under test to more speech instances, until the vocal expert has enough information to make his decision. This extension also falls outside the scope of this thesis.

Our choice between one of either basic architectures was not only based upon the descriptions presented above, but also on the results of the important research presented in [41]. In this paper, Viswanathan et al. have compared the serial and the parallel distributed decision fusion mechanisms. Their conclusions are:

1. For certain noise distributions, the parallel structure is not superior to the serial scheme. For additive white Gaussian noise (AWGN) and two sensors for instance, it can be shown that the serial fusion scheme performs better than the parallel one. However, with *three* or *more* sensors, the performance is essentially the same.
2. As a drawback, any serial network is vulnerable to link failures.
3. Considering the complexity of the serial scheme and the results from the (limited) comparative study, the choice seems to favor the parallel fusion for the distributed decision fusion problem.

The results of this study are a confirmation of the conclusions of the research presented in [36].

Taking into account the descriptions of the basic architectures and the results of the two studies mentioned above, we do opt for a parallel decision fusion scheme in the case of our application.

In a verification system with d experts in parallel, the decision fusion module using a binary decision scheme has to realize a mapping from the unitary hypercube of

$\mathbb{R}^d$ into the set {rejected; accepted}. A classifier having a d-dimensional input vector and two classes {rejected}, {accepted} is characterized by such a mapping. The *multi-expert* fusion module can therefore be considered as a *multi-dimensional* classifier. This particular classification case will be our standard fusion approach, since it allows to fall back immediately onto techniques available in the vast field of pattern recognition.

6. Paradigms

6.1. Overview

To solve our particular classification problem, classical parametric and non-parametric statistical pattern recognition techniques have been adapted. In the parametric class, piece-wise linear classifiers [47], and classifiers based on the general Bayesian decision theory (maximum a-posteriori probability and maximum likelihood), and on a simplified version of it (the Naive Bayesian classifier, which has been applied in the case of simple Gaussians and in the case of a logistic regression model), were experimented [46]. Furthermore experiments have been done using linear, quadratic, and multi-layer perceptron classifiers [45]. Also, several non-parametric classifiers have been used in this work [44]. A first representative has been the family of the (very simple: AND, OR) voting methods. Thereafter the k-NN classifier and some of its possible variants have been studied. And to finish with this class, another popular paradigm, the binary decision tree, has also been used.

6.2. Comments on the use of a bayesian framework

The main advantage of the Bayesian approach is that it leads to the optimal classifier, in the sense that it implements the lowest Bayes risk. There are however a number of problems with this approach. The most important problem is that the probability density functions (pdfs) have to be estimated correctly. This usually implies the selection of the structure (class of functions) for the approximator and the optimization of the free parameters to best fit the pdf. This optimization is performed on a training set. According to Occam's razor principle (which pleads for preferring the simplest hypothesis that fits the data [29]), the plasticity of the approximator has to be chosen carefully. For highly plastic approximators, quite general pdfs may be approached, but an important (often impossible to obtain) number of samples is needed for performing the training. Furthermore, the training set should be representative (which in general does not correspond to the equal a priori probability hypothesis) and over-training has to be avoided to reach good generalization [7]. On the other hand, by using an approximator with limited plasticity (few parameters, regularization techniques, etc.), fewer examples are needed but more a priori knowledge is intrinsically encoded by limiting the possible solutions. Poor a priori knowledge will lead to bad results. In practice, the best compromise should be searched, but the true MAP or ML decision rules can most of the time not be implemented and the theoretical minimal Bayes risk remains an unachievable lower bound which has as a consequence that in the field of pattern recognition and related disciplines, it is common practice to see that other, non-Bayesian, methods are being used. In this context it is worth mentioning Vapnik's result that it is easier – in an information theoretic sense – to estimate a classifier directly from data than estimating a distribution [43]. However, sometimes it is possible to justify some of those approaches in the light of the general Bayesian approach, which has the advantage of expliciting the underlying conditions/constraints. The first step towards deriving such specific cases is to introduce the classical hypothesis of independence, which leads to the so-called *naive Bayes classifier* [29]. This classifier has been used in our previous work, where we did assume that the probability distributions involved were: (1) simple Gaussian distributions, and (2) members of the exponential family with equal dispersion parameters (the logistic regression model).

7. Experimental comparison of classifiers

7.1. Test results

Table 7 gives an overall view of the best verification results obtained with the classifiers presented in this paper.

The first and most important observation we can make when looking at the results obtained by the fusion methods that we have experimented is that, in our application, fusion always improves the system performances beyond those of even the best single expert. The second observation is that these results seem to indicate that, generally speaking and again in our application, the class of the parametric methods does perform better than the class of non-parametric methods. Two indications that this statement is true in our case are that:

1. The mean TER calculated over the six parametric methods (1.10) is smaller than the mean TER calculated over the six non-parametric methods (1.95).
2. The mean rank (1 being attributed to the 'best' method (i.e. the method with the lowest TER), 2 to the 'second best', and so on) calculated over the six parametric methods (5.83) is smaller than the mean rank calculated over the six non-parametric methods (7.17).

Table 7
Summary table of verification results for the following fusion modules: maximum likelihood (ML), maximum a posteriori probability (MAP), logistic regression (LR), quadratic classifier (QC), linear classifier (LC), AND-voting rule (AND), OR-voting rule (OR), k-nearest neighbor (k-NN), k-nearest neighbor using vector quantization ($k-\mathrm{NN}+\mathrm{VQ}$), binary decision tree (BDT)

Method	FRR (%) (37 tests)	FAR (%) (1332 tests)	TER (%) (1369 tests)
ML	2.7 [0.5, 13.8]	0.7 [0.4, 1.3]	0.7 [0.4, 1.3]
MAP	5.4 [1.5, 17.7]	0.0 [0.0, 0.3]	0.1 [0.0, 0.5]
LR	2.7 [0.5, 13.8]	0.0 [0.0, 0.3]	0.1 [0.0, 0.5]
QC	0.0 [0.0, 9.4]	2.4 [1.7, 3.4]	2.3 [1.6, 3.2]
LC	0.0 [0.0, 9.4]	3.1 [2.3, 4.2]	3.0 [2.2, 4.0]
MLP	0.0 [0.0, 9.4]	0.4 [0.2, 0.9]	0.4 [0.2, 0.9]
OR	0.0 [0.0, 9.4]	7.4 [6.1, 8.9]	7.2 [5.9, 8.7]
AND	8.1 [2.8, 21.3]	0.0 [0.0, 0.3]	0.2 [0.1, 0.6]
MAJ	0.0 [0.0, 9.4]	3.2 [2.4, 4.3]	3.1 [2.3, 4.2]
$k-\mathrm{NN}$	8.1 [2.8, 21.3]	0.0 [0.0, 0.3]	0.2 [0.1, 0.6]
$k-\mathrm{NN}+\mathrm{VQ}$	0.0 [0.0, 9.4]	0.5 [0.2, 1.0]	0.5 [0.2, 1.0]
BDT	8.1 [2.8, 21.3]	0.3 [0.1, 0.8]	0.5 [0.2, 1.0]

From a first, intuitive, analysis it would seem like we find here three groups of methods. A first group with TER values lying between 0.1 and 0.7, a second group with values between 2.3 and 3.1, and finally the 'OR'-voting method with a TER result of 7.2. More specifically, the logistic regression method (a parametric method) gives the overall best TER results, and the OR-voting scheme (a simple non-parametric method) gives the overall worst TER results. To verify the good results obtained with the logistic regression model, we did a *validation* test using this method.

7.2. Validation results

These validation results have been obtained on the same M2VTS database, but this time using a more sophisticated protocol: the so-called *leave-one-out* method [16]. In this case the M2VTS database has been split in two groups: group 1 consisting of 18 persons and group 2 containing 19 persons. These two groups have been used in turn, respectively, as training and testing data set for the fusion module, in such a way that if one group was used for training, the other one was used for testing. The purpose of this is split is to introduce a total separation between the training and the testing data sets. The fact that therefore not only the impostors, but also the clients are different in the training and the testing data sets, has as a direct consequence that the use of individual thresholds is not possible. For each group, client and impostor accesses are generated, rotating through the first four shots of the database. For group 1 this leads to 4 × 18 = 72 client and 4 × 18 × 17 = 1.224 impostor accesses. For group 2 the same method leads to 4 × 19 = 76 client and 4 × 19 × 18 = 1.368 impostor accesses. So this strategy produces in total 148 client and 2.592 impostor tests. This validation test protocol is visualized in Fig. 13, and it is the same as the one described in [32].

The results obtained using this leave-one-out validation protocol are given in Table 8. This validation

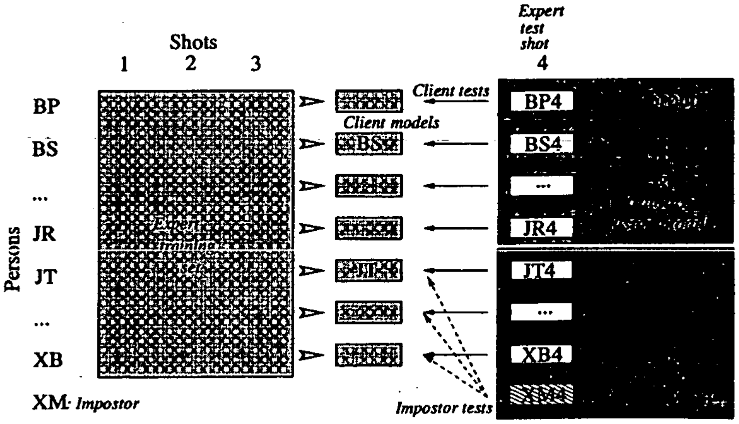

Fig. 13. Visualization of the leave-one-out validation protocol.

Table 8
Validation of the logistic regression using a leave-one-out protocol on the M2VTS database

Method	FRR (%) (148 tests)	FAR (%) (2.592 tests)	TER (%) (2.740 tests)
LR	0.0 [0.0, 2.5]	0.0 [0.0, 0.2]	0.0 [0.0, 0.1]

experiment shows indeed that the logistic regression does perform as well as predicted by the tests done using the original, more limited, test protocol. The verification performances obtained on the validation set extracted from this small database are extremely good, but when trying to generalize one should keep in mind the limitations of the described work.

7.3. Statistical significance

As can be observed in Table 7, the FAR, FRR and TER results from most of the methods used are lying very close to one another and they have confidence intervals which are overlapping each other in almost all cases. This means that, based on such FAR, FRR or TER results, it is not easy to decide without hesitation which method to use. That is why it would be very useful to have a systematic method which detects statistical significant differences between the FAR, FRR and/or TER results obtained by all the methods used.

In the case the scores obtained by the different fusion modules are independently drawn from normally distributed populations with the same variance, this problem can be solved by performing a basic ANOVA, supplemented with so-called ad hoc tests [30]. The ANOVA tells us *if* there are statistical significant differences between the different methods, and the ad hoc tests (least significant difference (LSD), Duncan's Multiple Range Test and many others) tells us *where* the statistical significant differences exactly are. It is reminded here that the statistical comparison needs to be done between all the methods at the same time (as opposed to a series of pairs-wise comparisons), to avoid the in the statistical community well-known effect of dramatically increasing the type one error [26,38].

Applying the ANOVA method in our case leads to a firm rejection of the hypothesis H0, which means that there are significant differences between the presented fusion modules. To find out where these statistical significant differences are, we did use Duncan's Multiple Range Test. The result of this ad hoc test with the highest power (i.e. the method with the lowest error of type II) is that three different groups of methods are significantly distinct. These groups are the following ones:

1. the first group (the one with the best performances) is formed by the following methods: LR, MAP, AND, k-NN, MLP, $k-\text{NN}+\text{VQ}$, BDT, and ML;
2. the second group consists of: QC, LC, and MAJ;
3. and finally the worst method in our case is the OR-vote.

However, since some of the fusion modules we use generate hard binary decisions as their output, we are, strictly speaking, not allowed to perform an ANOVA. Therefore we do have to use non-parametric methods. For the same reason as in the parametric case, it is here again absolutely necessary to compare all methods *at the same time*. In [18], five approximate statistical tests are presented for determining whether one learning algorithm out-performs another one on a particular learning task, but unfortunately these tests only allow a two-by-two comparison.

We believe that, depending on the discrete or continuous aspect of the output of the fusion module, two different non-parametric statistical tests can be applied to test the hypothesis H0: all used fusion methods are of equal performance, against the alternative H1: there are variations in performance. If one or more of the outputs of the fusion modules are *binary* or *hard* decisions, then *Cochran's Q test for binary responses* is suitable to solve this problem. If however the outputs of *all* fusion module are a continuous or *soft* decisions, then *Page's test for ordered alternatives* could be used. The latter test has the advantage that it has more *power* than the former [38,39].

Since in this specific case there are several fusion modules with binary (in case 0 for a reject and 1 for an acceptance) outputs, the only possibility is to use Cochran's Q-test. In conventional terms of this test, the different fusion methods are called the *treatments* and the different access tests are called the *blocks*. If we have t treatments and b blocks with binary responses, the appropriate test statistic is

$$Q = \frac{t(t-1)\sum_i T_i^2 - (t-1)N^2}{tN - \sum_j B_j^2},$$

where T_i is the total of 0s and 1s for treatment i, B_j the total for block j and N is the grand total. The exact distribution of Q is difficult to obtain, but for large samples Q has approximately a chi-squared distribution with $t-1$ df [39].

The application of Cochran's test for binary responses gives in our case a Q value which is much larger than the corresponding critical value of the corresponding chi-squared distribution, which leads us to reject the hypothesis H0. This means that there are significant differences between the presented fusion methods, which is the same conclusion as the one obtained by performing the ANOVA. And although Cochran's test does not say where exactly these differences are, we did however establish one thing for sure: the best method (which in our case is the logistic regression) is *statistically significant better* than the worst method (which in our case is the OR voting scheme). This result is rather trivial, since for these two 'extreme' methods the 95% confidence intervals do not overlap at all.

To conclude this section it can be seen that the results of the ANOVA and ad hoc tests (although strictly speaking not allowed) do reinforce our first, intuitive approach. The allowed Cochran's Q test does confirm the results of the ANOVA, but unfortunately we did not find a non-parametric equivalent for the ad hoc tests. Combining all this information leads us to the conclusion that there is no statistically justified evidence to prefer one specific method from the first (best performing) group above another one from the same group.

In our case however, we do have a strong preference for the logistic regression, based on the following considerations:

1. logistic regression did obtain the least number of errors (one single error on 1369 access tests) on the M2VTS database;
2. logistic regression uses the *soft decision* scores of the different experts, which do contain more information than just the *binary hard decision*;
3. logistic regression is the parametric method that needs the smallest number of coefficients to be estimated. The fact that this is a good property can be justified by a combination of the 'simplicity favoring' idea of Occam's razor principle [29] and of Ljung's observation that, in practice, the role of (model) identification is more often that of finding an *approximate* description, catching *some relevant* features, than that of determining the true, exact dynamics [25].

8. Conclusions

One of the most important conclusions after observing the results in Table 7 is that in our particular application fusion always improves the system performances beyond those of even the best single expert. Although the presented multi-modal verification systems based on logistic regression seems to perform almost perfectly, these results have to be seen in their correct perspective, taking into account the very limited database. In any case, these performances are much better than those of verification systems using only one of the presented modalities. The question of which fusion method should be chosen, is a much more difficult one to answer. A lot depends on the application. To be able to choose a number of potentially powerful fusion paradigms, it is important to have a (large) representative database of your application, which can be used for training purposes. It also helps if one is able to visualize the different populations (clients versus impostors), because in that specific case the choice of the fusion methods could be guided by the shape of the separation frontier between the two populations one wants to obtain.

References

[1] S. Ben-Yacoub, Multi-Modal Data Fusion for Person Authentication using SVM. IDIAP-RR 7, IDIAP, 1998.

[2] E. Bigün, J. Bigün, B. Duc, S. Fisher, Expert conciliation for multi-modal person authentication systems by bayesian statistics, in: Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, Crans-Montana, Switzerland, March 1997, pp. 327–334.

[3] J. Bigün, G. Chollet, G. Borgefors (Eds.), Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication, Crans-Montana, Switzerland, March 1997.

[4] F. Bimbot, G. Chollet, Assessment of speaker verification systems, in: Handbook of Standards and Resources for Spoken Language Systems, Mouton de Gruyter, 1997.

[5] F. Bimbot, I. Magrin-Chagnolleau, L. Mathan, Second-order statistical measures for text-independent speaker identification, Speech Commun. 17 (1–2) (1995) 177–192.

[6] F. Bimbot, L. Mathan, Second-order statistical measures for text-independent speaker identification, in: ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, April 1994.

[7] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford UK, 1995.

[8] G. Borgefors, Hierarchical chamfer matching: A parametric edge matching algorithm, IEEE Trans. Pattern Anal. Machine Intell. 10 (6) (1988) 849–865.

[9] R. Brunelli, D. Falavigna, Person identification using multiple cues, IEEE Trans. Pattern Anal. Machine Intell. 17 (10) (1995) 955–966.

[10] R. Chellapa, L. Davis, P. Phillips (Eds.), Proceedings of the Second International Conference on Audio- and Video-based Biometric Person Authentication, Washington DC, USA, March 1999.

[11] C.C. Chibelushi, J.S. Mason, F. Deravi, Integration of acoustic and visual speech for speaker recognition, EUROSPEECH '93, 1993, pp. 157–160.

[12] G. Chollet, C. Montacie, Evaluating speech recognizers and data bases, in: H. Niemann, M. Lang, G. Sagerer (Eds.), Recent Advances in Speech Understanding and Dialog Systems, NATO ASI F: Computer and Systems Sciences, vol. 46, Springer, Berlin, 1988, pp. 345–348.

[13] T. Choudhury, B. Clarkson, T. Jebara, A. Pentland, Multimodal person recognition using unconstrained audio and video, in: Proceedings of the Second International Conference on Audio- and Video-based Biometric Person Authentication, Washington DC, USA, March 1999, pp. 176–181.

[14] K. Choukri, Quelques approches pour l'adaptation aux locuteurs en reconnaissance automatique de la parole, Ph.D. Thesis, ENST, Paris, November 1988.

[15] B.V. Dasarathy, Decision Fusion, IEEE Computer Society Press, Silver Spring, MD, 1994.

[16] P.A. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[17] U. Dieckmann, P. Plankensteiner, T. Wagner, Sesam: A biometric person identification system using sensor fusion, Pattern Recog. Lett. 18 (9) (1997) 827–833.

[18] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput. 10 (7) (1998).

[19] B. Duc, G. Maître, S. Fischer, J. Bigün, Person authentication by fusing face and speech information, in: Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication, Lecture Notes in Computer Science, Springer, Berlin, 1997.

[20] L. Hong, A. Jain, Integrating faces and fingerprints for personal identification, IEEE Trans. Pattern Anal. Machine Intell. 20 (12) (1998) 1295–1307.

[21] A. Jain, R. Bolle, S. Pankanti. BIOMETRICS: Personal Identification in Networked Society, Kluwer Academic Publishing, Dordrecht, 1999, pp. 1–41 (Chapter – Introduction to biometrics).

[22] P. Jourlin, J. Lüttin, D. Genoud, H. Wassner, Acoustic–labial speaker verification, in: Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication, Lecture Notes in Computer Science, Springer, Berlin, 1997.

[23] G.K. Kanji, 100 Statistical Tests, Sage, Beverley Hills, CA, 1993.

[24] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Machine Intell. 20 (3) (1998) 226–239.

[25] L. Ljung, Convergence analysis of parametric identification methods, IEEE Trans. Automatic Control 23 (5) (1978) 770–783.

[26] B.F.J. Manly, Multivariate Statistical Methods, second ed., Chapman & Hall, London, 1994.

[27] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET curve in assessment of detection task performance, in: Eurospeech '97, Rhodes, Greece, 1997, pp. 1895–1898.

[28] J. Matas, K. Jonsson, J. Kittler, Fast face localization and verification, in: A. Clark (Ed.), British Machine Vision Conference, BMVA Press, 1997, pp. 152–161.

[29] T.M. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.

[30] D.C. Montgomery, Design and Analysis of Experiments, fourth ed., Wiley, Chichester, 1997.

[31] J. Oglesby, What's in a number? Moving beyond the equal error rate, in: ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, April 1994, pp. 87–90.

[32] S. Pigeon, Authentification multimodale d'identité, Ph.D. Thesis, Université Catholique de Louvain, February 1999.

[33] S. Pigeon, L. Vandendorpe, The M2VTS database (release 1.00), http://www.tele.ucl.ac.be/M2VTS, 1996.

[34] S. Pigeon, L. Vandendorpe, Profile authentication using a Chamfer matching algorithm, in: Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication, Lecture Notes in Computer Science, Springer, Berlin, 1997, pp. 185–192.

[35] M.A. Przybocki, A.F. Martin, NIST speaker recognition evaluations, in: Proceedings of the First International Conference on Language Resources and Evaluation, vol. 1, Granada, Spain, May 1998, pp. 331–335, ELRA.

[36] A. Reibman, L. Nolte, Design and performance comparison of distributed detection networks, IEEE Trans. Aerospace Electronic Systems 23 (6) (1987) 789–797.

[37] G. Saporta, Probabilités, analyse des données et statistique, vol. I, Editions Technip, 1990.

[38] S. Siegel, N.J. Castellan, Nonparametric Statistics, McGraw-Hill, New York, 1988.

[39] P. Sprent, Applied Nonparametric Statistical Methods, Chapman & Hall, London, 1989.

[40] SPSS. http://www.spss.com, 1998.

[41] S. Thomopoulos, R. Viswanathan, R. Tumuluri, Optimal serial distributed decision fusion, IEEE Trans. Aerospace Electronic Systems 24 (4) (1988) 366–376.

[42] H.L. Van Trees, Detection, Estimation and Modulation Theory, vol. 1, Wiley, New York, 1968.

[43] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, USA, 1998.

[44] P. Verlinde, G. Chollet, Comparing decision fusion paradigms using *k*-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application, in: Proceedings of the Second International Conference on Audio- and Video-based Biometric Person Authentication, Washington DC, USA, March 1999, pp. 188–193.

[45] P. Verlinde, P. Druyts, G. Chollet, M. Acheroy, A multi-level data fusion approach for gradually upgrading the performances of identity verification systems, in: Sensor Fusion: Architectures, Algorithms, and Applications III, vol. 3719, SPIE Press, Orlando, USA, April 1999.

[46] P. Verlinde, P. Druyts, G. Chollet, M. Acheroy, Applying Bayes-based classifiers for decision fusion in a multi-modal identity verification system, in: International Symposium on Pattern Recognition in Memoriam Pierre Devijver, Brussels, Belgium, February 1999.

[47] P. Verlinde, G. Maître, E. Mayoraz, Decision fusion using a multi-linear classifier, in: Proceedings of the International Conference on Multisource–Multisensor Information Fusion, vol. 1, Las Vegas, USA, July 1998, pp. 47–53.

[48] J.L. Wayman, BIOMETRICS: Personal Identification in Networked Society, Kluwer Academic Publishers, Dordrecht, 1999, pp. 345–368 (Chapter – Technical testing and evaluation of biometric identification devices).